

**Formation**

# **Plan de Gestion des Données (PGD)**

Sandrine Auzoux    Pauline Corbière  
Laurence Dedieu    Sébastien Paradis

**5 et 6 avril 2018**

# Programme de la formation

## Jour 1

Accueil, présentation des participants

1. Définitions clés en matière de données et PGD et cadre juridique général
2. Les enjeux du PGD
3. Construire le PGD
4. Décrire ses données
5. Sécuriser, stocker et archiver ses données

## Jour 2 (matin)

6. Bonnes pratiques juridiques
  7. Partage et valorisation des données
- Synthèse et évaluation

# DÉFINITIONS

# 1. Définitions clés

## → Donnée – Information – Connaissance

### Donnée

Résultat direct d'une mesure (faits, observations, éléments bruts)

*Ex. : 40°C*

### Information

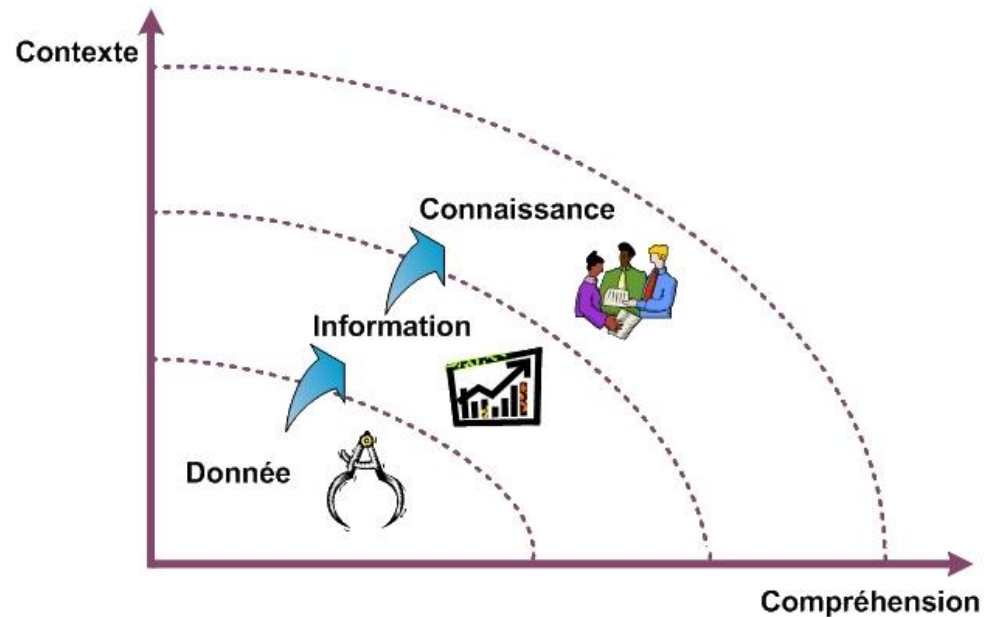
Donnée interprétée  
(qui, quoi, quand, où)

*Ex. : température de l'air en degré  
Celsius à 14h à Montpellier*

### Connaissance

Information comprise (comment, pourquoi)

*Ex. : il fait chaud. C'est l'été.*



# 1. Définitions clés

## → Nature et granularité des données

### Donnée brute

Donnée élémentaire issue d'une mesure, qui n'a encore été **ni organisée** dans une banque de données, **ni analysée, ni interprétée**.

*Ex. : index de compteur, hauteur d'eau au droit d'une station de jaugeage, etc.*

### Donnée corrigée (traitée/exploitable)

Donnée produite après formatage ou correction de données brutes.

### Donnée dérivée (ou analysée ou élaborée)

Donnée issue du traitement, de la combinaison ou de la réorganisation de données brutes, pour la rendre plus lisible ou la présenter sous une forme canonique.

*Ex. : un taux de collecte, le rendement moyen d'un essai.*

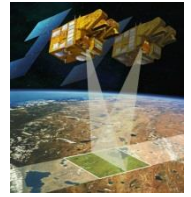


# 1. Définitions clés

## → Les données de la recherche

### Données d'observation

Données capturées en temps réel généralement uniques et impossibles à reproduire. Elles ont vocation à être conservées de façon pérenne.



télédétection



enquêtes

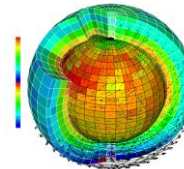
### Données expérimentales

Données obtenues à partir d'équipements en laboratoire, suivant une méthodologie bien définie. Potentiellement reproductibles, mais à des coûts parfois prohibitifs.

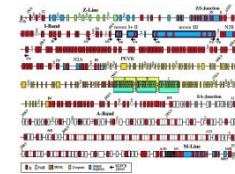


### Données de simulation

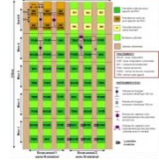
Données générées à partir de modèles.



Modèle climatique



Séquence de gènes



Résultats  
agronomiques

### Données dérivées ou compilées

Données résultant du traitement ou de la combinaison de données «brutes».



GenBank



Data mining – text mining

### Données de référence

Collection de jeux de données qui ont été revus, annotés et mis à disposition par les pairs.



# 1. Définitions clés

## → Les métadonnées

Les métadonnées sont des « *données qui décrivent des données* » :

- ❖ **Information** structurée associée à un "objet", un document ou un jeu de données
- ❖ **Documentation** qui permet à l'utilisateur de comprendre, de comparer et d'échanger le contenu du jeu de données décrit

Il existe des **standards** de métadonnées :

- ❖ Standards minimaux (ex : Dublin Core)
- ❖ Standards métiers (ex : EML, DDI...)



Il est conseillé de produire les métadonnées au **moment de la collecte ou de la création** des données plutôt qu'à posteriori. Les métadonnées seront **complétées tout au long du cycle de vie** des données.



*Un objet sans étiquette n'est connu que de son auteur*

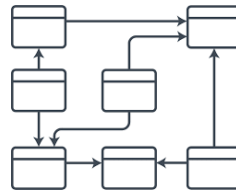


# 1. Définitions clés

## → Les bases de données

Une base de données est un **ensemble de données** stockées de façon :

- ❖ **Persistante** : stockage permanent
- ❖ De **redondance minimale** : la même information n'est idéalement présente qu'une fois (unicité)
- ❖ **Exhaustive** : la base de données contient toutes les informations requises pour le service attendu
- ❖ **Structurée** : la structure est définie dans un schéma (le « modèle »)



- ❖ Elle est gérée par un système de gestion de bases de données

**ORACLE**





# 1. Définitions clés

## → Les jeux de données

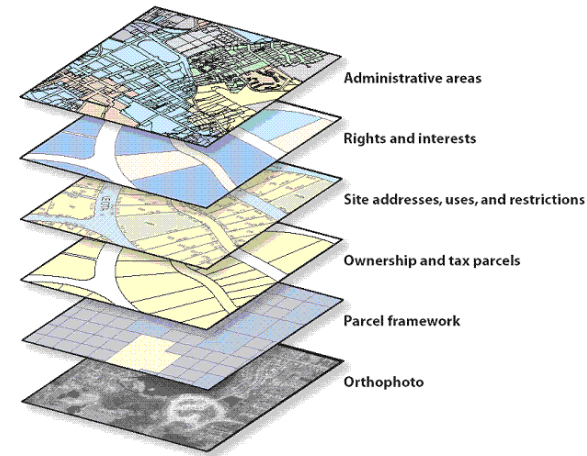
Un jeu de données (*dataset*) est l'agrégation d'enregistrements de données organisés pour former un **ensemble cohérent**. Les jeux de données numériques sont formatés de telle sorte qu'ils soient communicables, interprétables et adaptés à un traitement informatisé.



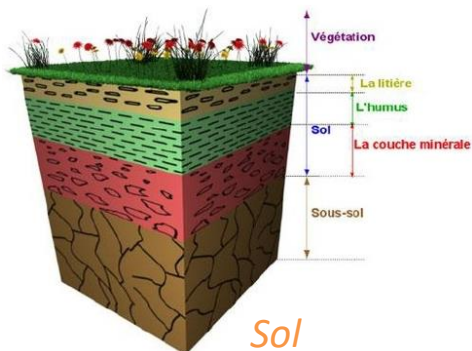
Météo



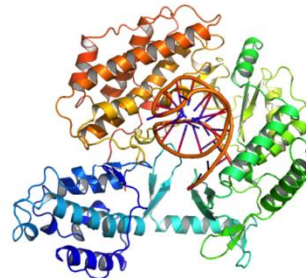
Cartographie d'acteurs



Couches SIG



Sol



Biologie cellulaire



Culture



Séries temporelles

# 1. Définitions clés

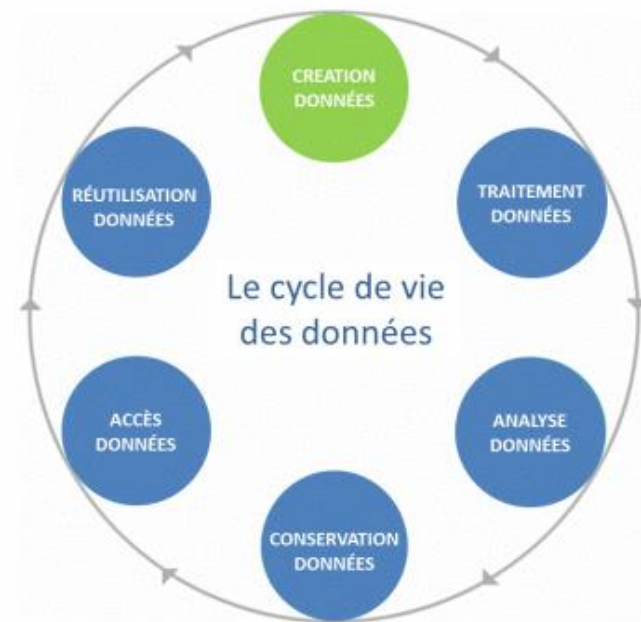
## → Le plan de gestion des données

Le PGD est un document qui :

- ❖ Rassemble les **règles de gestion** et de **documentation** des données produites et réutilisées au cours d'un projet de recherche
- ❖ Renseigne les modalités de **partage**, de **conservation** et de **valorisation** des données après la clôture du projet
- ❖ Favorise la **compréhension**, la **diffusion** et la **réutilisation** des données

Il doit être **mis à jour** tout au long du projet : 3 versions minimums en début, milieu et fin de projet.

Il s'appuie sur le **cycle de vie des données**.



# 1. Définitions du point de vue juridique

## → Les données

- ❖ Pas de définition légale de la donnée
- ❖ Arrêté du 22 décembre 1981 sur l'enrichissement de la langue française : représentation d'une information conventionnelle destinée à faciliter son traitement
- ❖ Définition de l'OCDE (Organisation de coopération et de développement économique) des données de la recherche :  
*Les données de recherche sont définies comme des enregistrements factuels (chiffres, textes, images et sons) qui sont utilisés comme sources principales pour la recherche scientifique et sont généralement reconnus par la communauté scientifique comme nécessaires pour valider des résultats de recherche. Un ensemble de données de recherche constitue une représentation systématique et partielle du sujet faisant l'objet de la recherche.*
- ❖ D'un point de vue juridique pas de différence entre données brutes, élaborées ou métadonnées

# 1. Définitions du point de vue juridique

## → Les bases de données

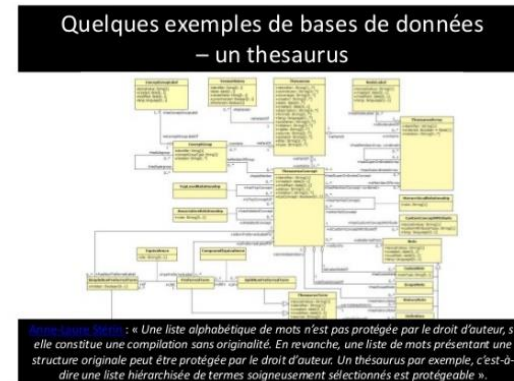
### Définition d'une base de données

La base de données est définie à l'article L112-3 du Code de la Propriété Intellectuelle :

« On entend par base de données un **recueil** d'œuvres, **de données** ou d'autres éléments indépendants, **disposés de manière systématique ou méthodique**, et **individuellement accessibles** par des **moyens électroniques** ou par **tout autre moyen**. »



*Cour d'Appel de Paris, 12 septembre 2001* : Peu importe que la base de données soit sous une forme papier, le support est indifférent.



# 1. Définitions clés du point de vue juridique

## Le producteur de base de données

Personne qui prend l'initiative et le risque des investissements correspondants. Le producteur de la base de données bénéficie d'une protection du contenu de la base lorsque la constitution, la vérification ou la présentation de celui ci atteste d'un investissement financier, matériel ou humain substantiel.

## Métadonnées

Données servant à définir ou décrire une autre donnée quel que soit son support (papier ou électronique). La métadonnée possède le même statut juridique que la donnée.

*Ex. : des coordonnées GPS*

## Jeu de données

Un jeu de données (*dataset*) peut être une base de données ou seulement des données.



# 1. Définitions du point de vue juridique

## → Introduction au cadre juridique des données et bases de données

### Le principe

Les données sont de libre parcours et ne sont pas appropriables.

### Les exceptions

1. Certaines données sont soumises au **droit d'auteur** : plans, photographies, croquis, etc....
  - ❖ Condition : critère d'originalité
  - ❖ Conséquences : pour utiliser ces données il faut une autorisation expresse écrite et préalable de l'auteur de la donnée protégée (exception : la courte citation).
2. Données organisées en base de données : **droit sui generis du producteur de base de données**.
3. La structure de la base de données peut bénéficier de la protection par le droit d'auteur.

# 1. Définitions clés du point de vue juridique

## Cadre juridique de la base de données

La base de données doit être appréhendée sous 3 angles :

- ❖ Les données contenues dans la base de données (le contenu)
- ❖ La structure de la base de données (le contenant)
- ❖ La base de données dans son ensemble (contenant + contenu)

# 1. Définitions clés du point de vue juridique

## La donnée

- ❖ Par principe les données ne sont pas appropriables
- ❖ « *La protection des bases de données par le droit d'auteur prévue par la présente directive ne couvre pas leur contenu et elle est sans préjudice des droits subsistant sur ledit contenant* »
- ❖ La donnée en elle-même ne fait pas l'objet de protection

### **Exceptions**

La donnée peut être protégée par un droit de propriété intellectuelle : droit d'auteur (dessin, photographie...), droit des marques : nécessité d'obtenir une licence pour exploiter ces données.

# 1. Définitions clés du point de vue juridique

## La structure

La structure ou architecture de la base de données peut faire l'objet d'une **protection par le droit d'auteur**.

- ❖ Condition : la structure doit être **originale**
- ❖ Pas évident de retrouver l'empreinte de la personnalité de l'auteur dans une base de données
- ❖ Critère du législateur : art.L122-3 : la structure est protégeable lorsque « par le choix ou la disposition des matières, elles constituent des créations immatérielles »
- ❖ Critère retenu par les juges : empreinte de la personnalité de l'auteur est caractérisée par les choix opérés pour mettre au point la structure ; parfois le critère de la nouveauté est retenu
- ❖ En pratique : considérer que toute structure est originale

# 1. Définitions clés du point de vue juridique





## La base de données dans son ensemble

- ❖ Le droit du producteur de la base de données est le **droit d'interdire l'extraction et/ou la réutilisation de tout ou partie du contenu de la base**
- ❖ Lutter efficacement contre le pillage des bases de données, à la portée de tous notamment avec le développement du numérique
- ❖ Protéger l'initiative et l'investissement réalisé pour mettre au point une base de données
- ❖ Transposition de la directive du 11 mars 1996 dans la loi 1<sup>er</sup> juillet 1998
- ❖ Articles L.341-1 et suivants du code de la propriété intellectuelle



# 1. Définitions clés du point de vue juridique

## La base de données

Date	Lieu	Taille	Photo	Climat	Température moyenne
2015	Brésil	240 cm		Humide	26°
2014	Afrique du Sud	610 cm		Sec	32°
2013	Togo	460 cm		Sec à très sec	29°
2012	Portugal	420 cm		Humide à sec	28°
2011	Côte d'Ivoire	800 cm		Sec	24°

# 1. Définitions clés du point de vue juridique

## La structure

Date	Lieu	Taille	Photo	Climat	Température moyenne
2015					
2014					
2013					
2012					
2011					

# 1. Définitions clés du point de vue juridique

## Les données

❖ 2015, 2014, 2013, 2012, 2011

❖ Brésil, Afrique, Togo, Portugal

❖ 240 cm, 610 cm, 460cm, 420 cm, 800 cm

❖ Photos



Pas de  
protection  
Données de  
libre parcours

Exception : Droit  
d'auteur sur les  
photographies

# 1. Définitions clés du point de vue juridique

## Titularité des droits

### Droit du producteur

Personne morale / employeur (celui qui prend le risque financier).

### Droit d'auteur sur la structure

Personne physique qui « écrit » la structure (chercheur, informaticien...).

### Données

Pas de droit sauf si la donnée bénéficie de la protection par le droit d'auteur = personne physique ayant réalisé l'œuvre (chercheur ayant réalisé la photographie).

# LES ENJEUX DU PGD



## 2. Les enjeux du PGD



### La réutilisation des données

- ❖ Réutilisation = démarche + engagée que partage ou accès  
→ documentation/contextualisation/mise en forme des données
- ❖ Minimiser les risques de perte de données
- ❖ Éviter les duplications d'expériences
- ❖ Optimiser le financement de la recherche
- ❖ Données = bien public
- ❖ Assurer transparence, intégrité, traçabilité, fiabilité
- ❖ Faire avancer la science et accélérer l'innovation

## 2. Les enjeux du PGD



Faciliter la réutilisation des données implique :

Mettre en œuvre des bonnes pratiques :

- ❖ De **description et documentation**,  
objectif d'interopérabilité des données  
(standards de métadonnées, vocabulaires contrôlés, normes, ...)  
en accord avec votre communauté scientifique
- ❖ De **gestion** des fichiers
- ❖ De **conservation** des données
- ❖ De **traçabilité des droits**

# 2. Les enjeux du PGD

Le PGD concerne tous producteurs de données

Doctorant

Scientifique :  
*Retraite  
changement d'activités*

Travail  
en équipe

Projet en  
partenariat

Facilite l'échange de données,  
la comparaison/combinaison de données  
Besoin de normes/métadonnées communes

Partage  
des données

Pour qui ?

Assurer la compréhension et réutilisation

Mise en œuvre de **Bonnes pratiques**



## 2. Les enjeux du PGD



Etes-vous sûrs de pouvoir **comprendre et retrouver vos** derniers jeux de **données** ?

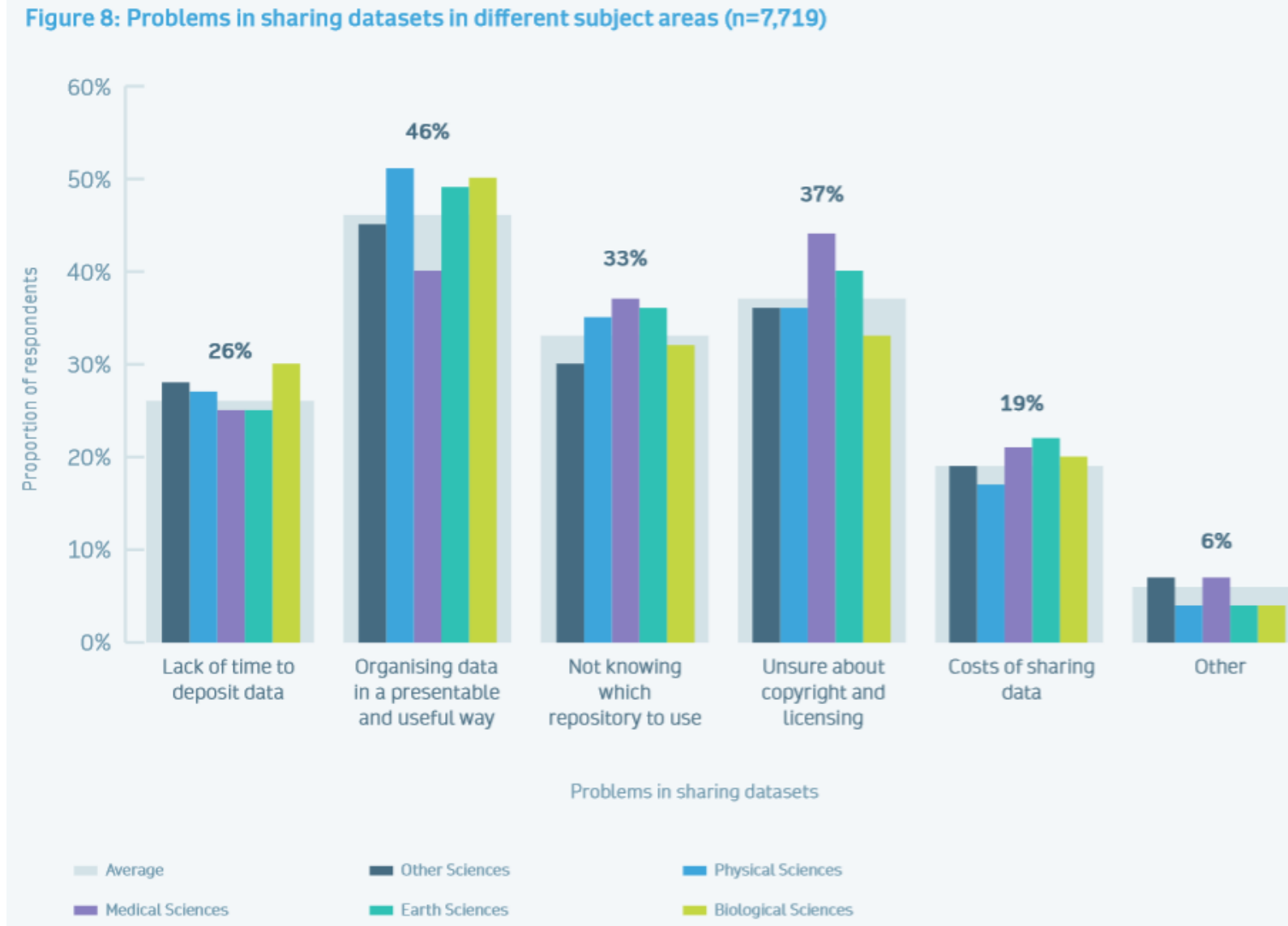
Par exemple :

- ❖ Ceux qui vous ont permis de publier un article de recherche
- ❖ Ceux que vous avez produits lors de votre dernier projet de recherche
- ❖ Ceux sur lesquels vous avez travaillé il y a 2 ans
- ❖ Ceux qui ont été produits par votre doctorant
- ❖ Etc.

→ Répondre au questionnaire (15 min)

## 2. Les enjeux du PGD

### → Barrières au partage des données





## 2. Les enjeux du PGD

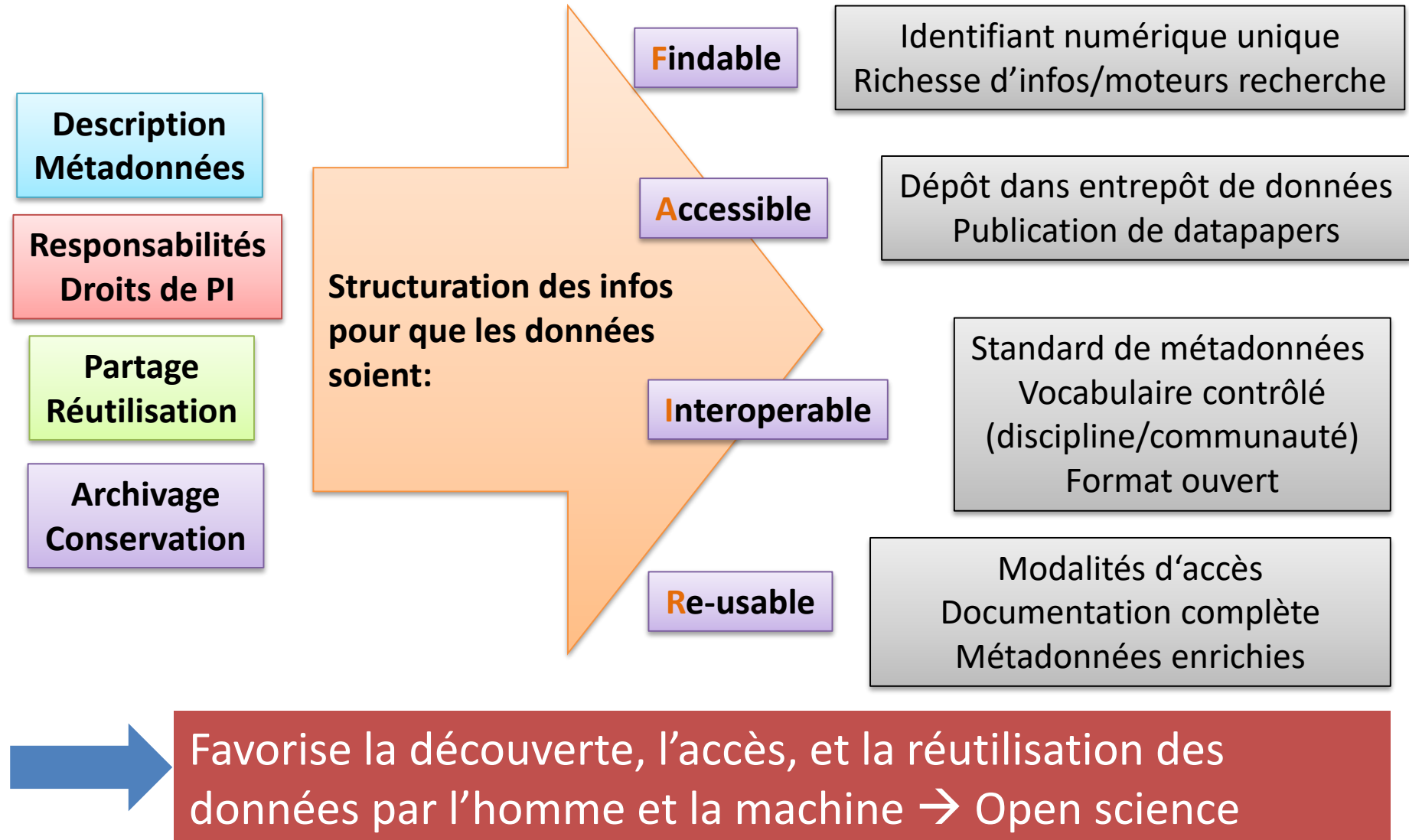


Pose de bonnes questions dès le début du projet

- ❖ Comment décrire au mieux les données pour assurer leur compréhension ?
- ❖ Quelles données conserver ? Comment ?
- ❖ Quelles données partager ? Avec qui ?
- ❖ Avons-nous les droits ?
- ❖ Aspects éthiques à prendre en compte ?  
*données sensibles, personnelles, du sud,  
ayant un impact sur homme/environnement*
- ❖ Combien ça va couter ?

# 2. Les enjeux du PGD

## Le PGD selon les principes FAIR

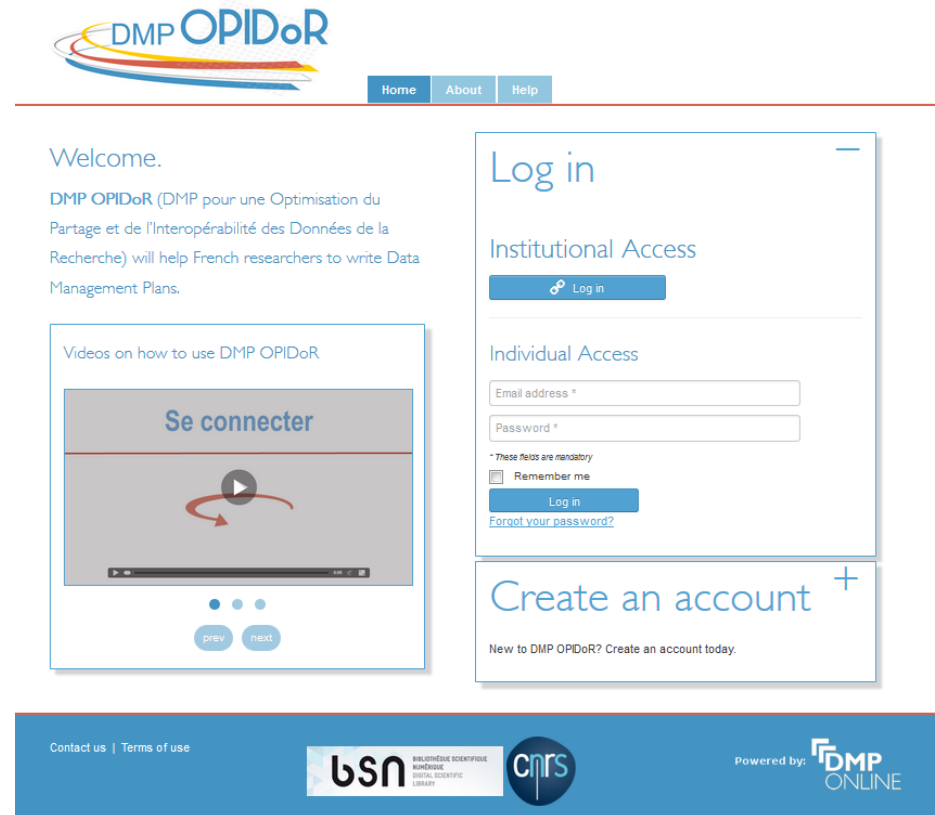


# 2. Les enjeux du PGD

## Pour rédiger le PGD



## Outil d'aide à la rédaction du PGD



Les 2 modèles, PGD classique et FAIR, sont disponibles

# 2. Les enjeux du PGD



## Bénéfice de la réutilisation de données

Si vous partagez vos données :

- ❖ Plus de visibilité
  - ❖ Plus de citations
  - ❖ Plus de notoriété
  - ❖ Nouvelles collaborations



Si vous avez accès à des données fiables  
qui remplacent 2 ans d'expérimentation :

- ❖ Gain de temps
  - ❖ Gain financier
  - ❖ Gain scientifique



## Pour le scientifique

- ❖ Mise en place de **bonnes pratiques** de gestion et documentation de ses données
- ❖ Réflexion sur les aspects **éthiques et juridiques**
- ❖ **Valorisation des données** : *datapaper*, stockage en entrepôt, mise en catalogue, etc.
- ❖ Renforcement de la **coopération** avec services d'appui : IST, DSI, juridique, valorisation...
- ❖ **Montée en compétence** en gestion des données et tout ce qui est en relation avec l'Open Science

#### Pour le collectif projet

- ❖ Offre une **vue globale** du projet et **flux de données**
- ❖ Favorise et facilite le **travail collaboratif**
- ❖ Outil de **support du management transversal**

#### Pour l'institution

- ❖ Participe à la mise en œuvre d'une **politique de partage** des données (**Open Science**)
- ❖ Constitution de **catalogues de données** réutilisables

## Pour le scientifique

- ❖ Nécessite du temps (réflexion, rédaction, ...)  
→ **Utiliser un modèle documenté, un outil d'aide à la rédaction**
- ❖ Crainte récurrente par rapport à l'Open Data  
→ **Un PGD n'oblige pas à la diffusion des données**  
→ **Conditions de réutilisation au choix de celui qui publie**
- ❖ Incompréhensions : « Comment décrire un jeu de données encore non produit ? »  
→ **Peu de métadonnées servent à décrire un jeu de données**
- ❖ Désintérêt : perçu comme une tâche administrative  
→ **« Outil de travail qui facilite la gestion et le partage des données tout au long du projet de recherche » (S. Pamerlon, GBIF)**



## Pour le collectif projet

- ❖ Définir la granularité des jeux de données
  - **Spécifique à chaque projet !**
  - **Pistes de réflexion : potentiel de réutilisation des données**
  
- ❖ Susciter l'intérêt des membres du projet...
  - **Accompagnement : assistance, formation, organisation d'ateliers...**
  
- ❖ ... tout au long du projet !
  - **Animation continue, mise en place d'outils collaboratifs**

3

# CONSTRUIRE LE PGD

# 3. Construire le PGD

Logged in as Sébastien PARADIS ▾

[View plans](#)[Create plan](#)[About](#)[Help](#)[Admin area](#)

## My plans

The table below lists the plans that you have created, and any that have been shared with you by others. These can be edited, shared, exported or deleted at anytime.

Filter plans <span>✕</span> <span>Filter</span> <span>⚙</span>				
Name ▾	Owner ▾	Shared? ▾	Last edited ▾	Select an action ▾
<a href="#">PGD ArchiWood</a>	Me	No	26-09-2016	<a href="#">Edit</a> <a href="#">Share</a> <a href="#">Export</a> <a href="#">Delete</a>
<a href="#">Statistiques Utilisation Alfresco</a>	Xavier Rouviere	Yes (with 1 people)	24-11-2016	<a href="#">Edit</a> <a href="#">Export</a>
<a href="#">BREEDCAFS (BREEDing Coffee for AgroForestry Systems)</a>	Laurence Dedieu	Yes (with 1 people)	11-05-2017	<a href="#">View</a> <a href="#">Export</a>
<a href="#">PGD test validation</a>	Me	No	12-03-2018	<a href="#">Edit</a> <a href="#">Share</a> <a href="#">Export</a> <a href="#">Delete</a>
<a href="#">Données du projet ArchiWood</a>	Me	No	21-03-2018	<a href="#">Edit</a> <a href="#">Share</a> <a href="#">Export</a> <a href="#">Delete</a>

[Create plan](#)[Contact us](#) | [Terms of use](#) |

Powered by: **DMP**  
ONLINE



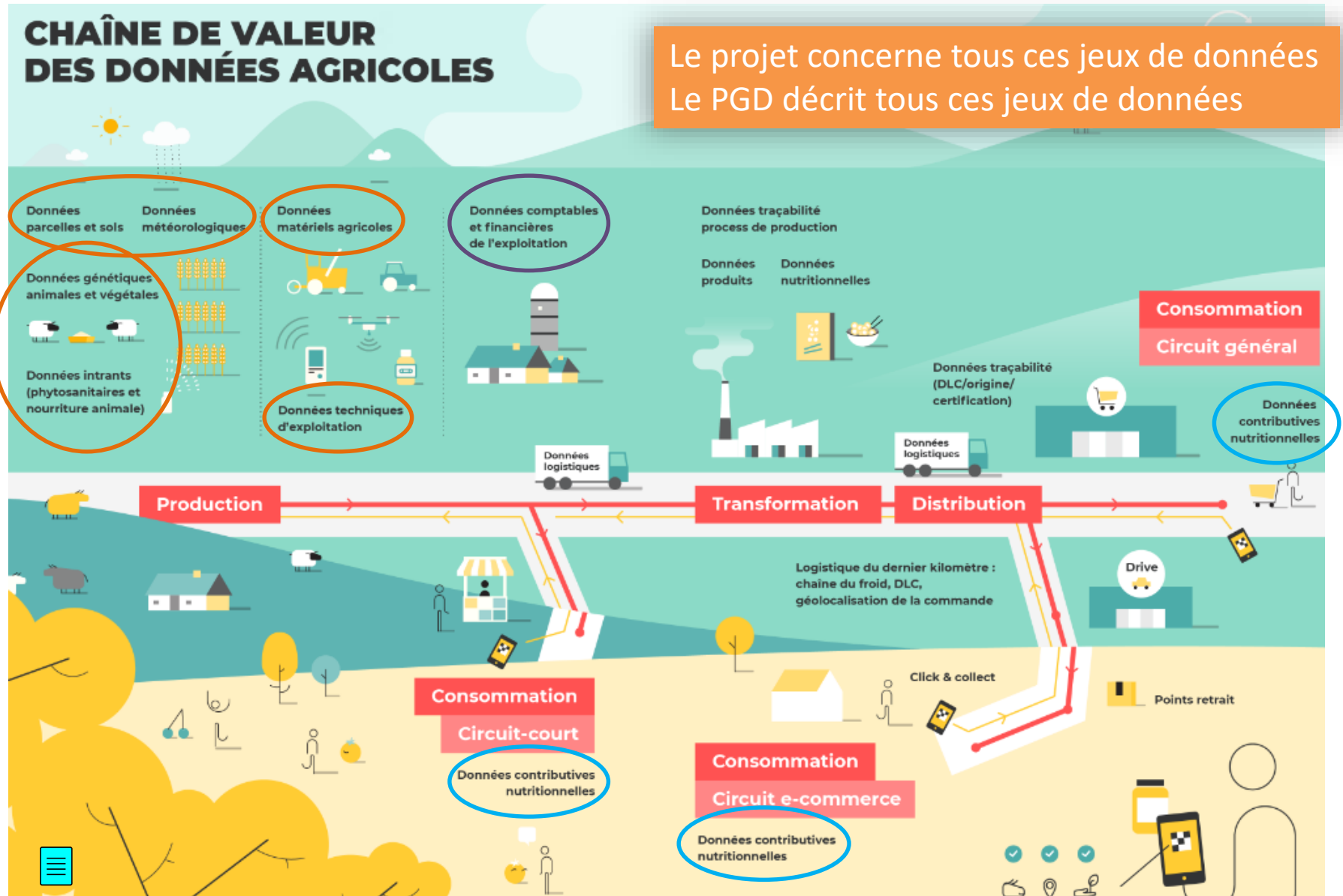
# DÉCRIRE SES DONNÉES

# 4. Décrire ses données



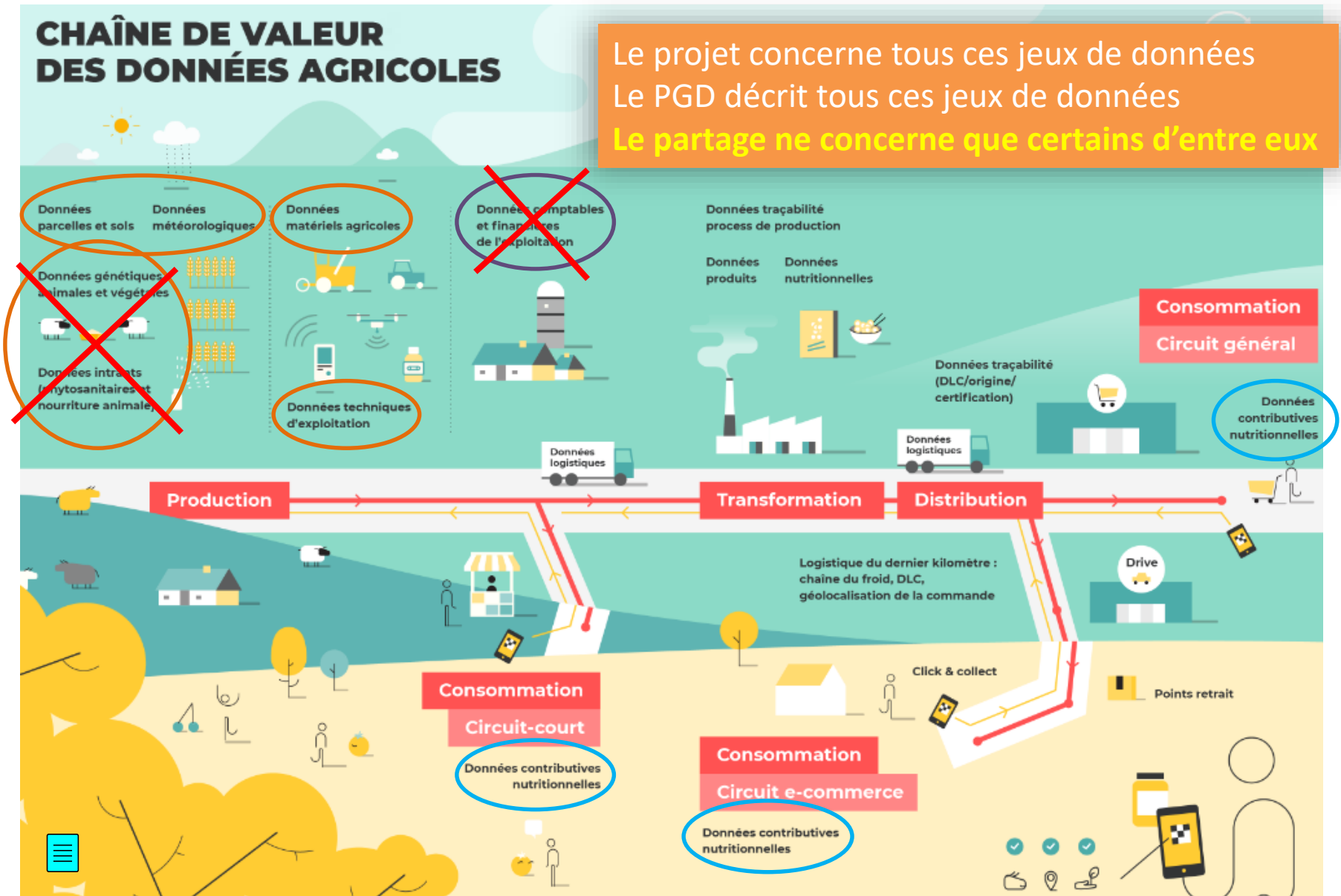
Le PGD doit décrire tous les jeux de données  
produits dans un projet  
Mais quelle granularité appliquer ?

## 4. Décrire ses données → Granularité des jeux de données



## 4. Décrire ses données

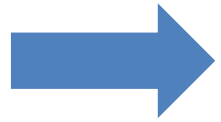
### → Granularité des jeux de données





## 4. Décrire ses données

### → Granularité des jeux de données

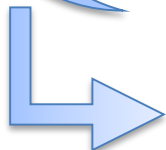


Projet BREEDCAFS (*BREEDing Coffee for AgroForestry Systems*)

Diversité des types de jeux de données :

- ❖ Physiologiques (racines, feuilles, tiges, fruits des caféiers)
  - phénotypiques
  - morphométriques
- ❖ Génétiques et épigénétiques
- ❖ Transcriptomiques
- ❖ Biochimiques
- ❖ Performance agronomique
- ❖ Sensorielles
- ❖ Enquêtes auprès des fermiers

5 pays/40 sites  
X hybrides + variétés



BREEDCAFS Database

## 4. Décrire ses données

### → Granularité des jeux de données

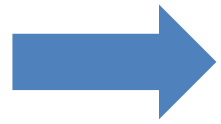


#### Projet BREEDCAFS : version 1 du PGD

- ❖ Présente l'ensemble des jeux de données produits  
Décrit le processus de création de la BD
- ❖ Choix du grain des jeux de données :
  - Par discipline** → cohérence technique  
même type/nature, méthodes/métadonnées
  - Par sites, pays, work package** → cohérence scientifique  
dépend du projet
- ❖ Prendre en compte :
  - La **valeur scientifique** et le **potentiel de réutilisation**
  - Le **devenir** : publications / entrepôt

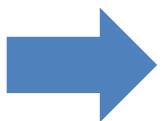
## 4. Décrire ses données

### → Granularité des jeux de données



#### Projet BREEDCAFS : version 1 du PGD

- ❖ Présente l'ensemble des jeux de données par discipline  
Décrit le processus de création de la BD
- ❖ Mentionne les jeux de données qui seront partagés :
  - Génétiques
  - Transcriptomiques
- ❖ Précise que pour les autres, il y aura concertation
  - Sensorielles (partenaire privé)
  - Enquêtes auprès des fermiers (anonymisation)
- ❖ Précise que la BD ne sera pas en accès libre



V2 prévue Nov 2018 :  
description améliorée des jeux de données attendue...

# 4. Décrire ses données

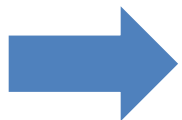
## ❖ Avec des métadonnées (éléments descriptifs structurés)

### ➤ ≠ standards de métadonnées

- Génériques
- disciplinaires

## ❖ Avec la documentation associée

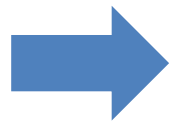
- Méthodes, protocoles, plan échantillonnage → fichier « Lisez moi »
- Dictionnaires des variables, Unités de mesure, abbréviations
- Equipement utilisé, méthode de calibration, contrôles
- Questionnaires, guides d'enquêteurs, technique d'anonymisation
- Schéma de la base de données, fichiers de syntaxe
- Liens vers les articles publiés à partir de ces données



Richesse des métadonnées + documentation complète  
→ assure la compréhension et réutilisation de vos données

## 4. Décrire ses données

### → Le standard de métadonnées génériques



#### Dublin Core : 15 métadonnées de base

DC Element	Notes
<b>Title</b>	Title of Data Collection
<b>Creator</b>	Authoring Entity of Data Collection
<b>Subject</b>	Keyword(s)
<b>Description</b>	Abstract
<b>Publisher</b>	Producer of Data Collection
<b>Contributor</b>	
<b>Date</b>	Production Date - Data Collection
<b>Type</b>	Kind of Data
<b>Format</b>	Type of File
<b>Identifier</b>	ID Number - Data Collection
<b>Source</b>	Sources - Used for Data Collection
<b>Language</b>	
<b>Relation</b>	Other Study Description Materials
<b>Coverage</b>	Time Period Covered
	Country
	Geographic Coverage
<b>Rights</b>	Copyright - Data Collection



Peu d'infos structurées  
décrivant les données

Habitat, climat ?  
Mode de culture ?  
Traitement ?  
Plante ? Organe ?  
Espèce ? Genre ?  
Souche ?

+ les métadonnées sont enrichies  
+ les données sont :  
trouvables  
compréhensibles  
réutilisables

# 4. Décrire ses données



Connaissez-vous des **standards de métadonnées** communément utilisés **dans votre domaine** ?



Connaissez-vous la **norme pour décrire une date** ou une période ?

## 4. Décrire ses données

### → Standards de métadonnées disciplinaires

- ❖ **Date / période / intervalles** : Norme ISO-8601 : aaaa-mm-jj
- ❖ **Géolocalisation**
  - Geospatial metadata standard, norme ISO 19115
  - Norme européenne INSPIRE
- ❖ **Ecologie, Biodiversité**
  - Ecological Metadata Language (EML)
  - Darwin Core
- ❖ **Sciences Humaines et Sociales**
  - Data Documentation Initiative (DDI)
- ❖ **Génomique** : MIxS standards
  - *Minimum Information about a Genome Sequence ; about a MetaGenome Sequence*
  - *Minimum Information about a MARKer gene Sequence; about a microarray*
  - déclinés pour 14 types d'environnement (Plant, soil, air, water, sediment,...)
- ❖ **Phénotypage de plantes**
  - MIAPPE: Minimum Information About a Plant Phenotyping Experiment



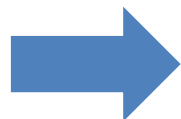
## 4. Décrire ses données

### → Darwin Core : standard en biodiversité

- ❖ **Classification**, ... , Vernacular name, ...
  - "class", "order", "family", "genus", "subgenus"
- ❖ **Basis of Record**
  - Preserved Specimen, Fossil Specimen, Living Specimen
  - Human Observation, Machine Observation
- ❖ **Sex**
  - "female", "hermaphrodite", "male"
- ❖ **lifeStage**
  - "egg", "eft", "juvenile", "adult", "2 adults 4 juveniles"
- ❖ **establishmentMeans**
  - "cultivated", "invasive", "escaped from captivity", "wild", "native"
- ❖ **habitat**
  - "oak savanna", "pre-cordilleran steppe"



**Freshwater Biodiversity Data Portal**  
The gateway to freshwater biodiversity data



DwC : permet de rassembler des millions de données d'occurrences d'espèces dans le portail GBIF

# 4. Décrire ses données

## → EML: Ecological Metadata Language

### Class II. Research origin descriptors

#### Site Description

- Site type
- Geography (location, size)
- Habitat
- Geology, Landform
- Climate

#### Experimental or sampling design

- Design characteristics
- Variables included
- Species sampled
- Data collection period, frequency

#### Research methods

- Field/Laboratory
- Instrumentation

SANParks

South African National Park Data Repository

The Knowledge Network for Biocomplexity

KNB

### Class III. Data set status and accessibility

#### Status

- Latest update
- Latest archive date
- Metadata status
- Data verification

#### Accessibility

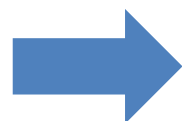
- Storage location and medium
- Contact person(s)
- Copyright restriction
- Costs



### Class IV. Data structural descriptors

#### Data Set Files

- Data set Identity
- Size
- Format and storage mode



EML : permet l'interopérabilité des données en écologie  
facilite la recherche, interprétation, sélection et exploitation

## 4. Décrire ses données

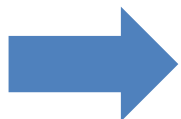
→ DDI (Data Documentation Initiative) : norme en SHS

### Description des données

- Unité d'observation
- Couverture géographique
- Couverture temporelle (AAAA-MM-JJ)
- Univers

### Méthodologie et traitement

- Dimension temporelle de l'enquête
- Collecteur des données
- Autres contributeurs
- Producteur
- Méthode d'échantillonnage
- Mode et Fréquence de collecte
- Opérations de contrôle
- Pondération
- Nettoyage des données



DDI : permet la réexploitation des données d'enquêtes grâce à la précision des données de contexte.

# 4. Décrire ses données

Données sur la composition chimique de feuilles de cacaoyers collectées en Côte d'Ivoire

	A	B	C	D	E	F	G	H	I	J
1	ID	Lon	Lat	region	Produc	N	P	K	Ca	Mg
2	Abe-1	-3.640017	6.712517	Abengourou	464	2,15	0,17	2,28	0,85	0,56
3	Abe-10	-3.695350	6.660200	Abengourou	763	2,21	0,13	2,94	1,56	0,56
4	Abe-2	-3.640050	6.710700	Abengourou	429	2	0,13	2,29	1,55	0,53
5	Abe-244	-3.248250	6.727933	Abengourou	600	2,21	0,14	2,6	1,11	0,48
6	Abe-245	-3.304450	6.653750	Abengourou	714	2	0,14	1,28	1,79	0,63
7	Abe-247	-3.300417	6.611333	Abengourou	625	1,89	0,15	2,69	1,53	0,72
8	Abe-268	-3.635183	6.568350	Abengourou	300	2,15	0,15	2,54	0,99	0,49
9	Abe-270	-3.659550	6.655417	Abengourou	650	1,78	0,17	3,03	0,67	0,56
10	Abe-271	-3.645900	6.673800	Abengourou	530	2,36	0,15	2,41	1,04	0,57
11	Abe-272	-3.701567	6.667383	Abengourou	355	2,05	0,19	2,2	1,16	0,71
12	Abe-3	-3.638017	6.706067	Abengourou	349	1,94	0,15	2,24	1,55	0,7
13	Abe-4	-3.645217	6.699017	Abengourou	256	2,1	0,15	2,72	1,32	0,57
14	Abe-5	-3.642550	6.687050	Abengourou	875	2,15	0,17	2,28	0,85	0,56

Plantations de cacaoyers adultes  
Echantillonnage en mars 2015

ID= Identifiant de chaque échantillon  
Localisation géographique  
Produc = yield of dry beans (kg/ha)  
N, P, K, Ca, Mg : %



Est-ce suffisant pour comprendre et réutiliser les données ?

## 4. Décrire ses données

→ Trouver un standard, une norme, un vocabulaire contrôlé



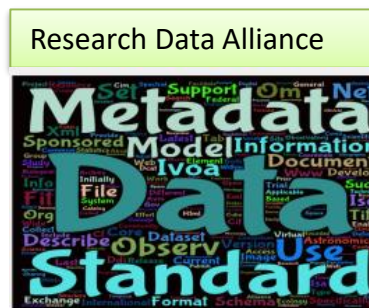
### Guides de standards de métadonnées



<https://fairsharing.org/>



<http://www.dcc.ac.uk/resources/metadata-standards>



<http://rd-alliance.github.io/metadata-directory/>

- ❖ Tenir compte de l'entrepôt de données ciblé et de la revue (si publication prévue)
- ❖ S'il y en a pas, expliquer comment vous créez les métadonnées ? et présentez les outils que vous utilisez ?

## 4. Décrire ses données

### → Qualité des données

→ S'assurer que les données soient cohérentes et correctes, en termes de précision, conformité, disponibilité...

Les principes de qualité des données peuvent s'appliquer à toutes les étapes :

- ❖ Saisie et enregistrement au moment de la collecte
- ❖ Manipulation préalable à leur numérisation (la préparation des étiquettes, l'inscription des données dans un registre, etc.)
- ❖ Identification de la collecte (spécimen ou observation) et son enregistrement
- ❖ Numérisation
- ❖ Documentation (saisie des métadonnées)
- ❖ Stockage et archivage
- ❖ Présentation et diffusion (publication, site web, etc.)
- ❖ Utilisation (analyse et manipulation)

## 4. Décrire ses données

### → Qualité des données

#### Recommandations de mise en œuvre

- ❖ Renseigner les valeurs d'incertitude(s)
- ❖ Documenter le processus de validation / nettoyage (nommer le(s) expert(s))
- ❖ Fournir des descriptions appropriées aux données (métadonnées)
- ❖ Utiliser des vocabulaires standards pour les métadonnées et les données quand ils existent
- ❖ Spécifier la licence sous laquelle les données peuvent être réutilisées
- ❖ Respecter les exigences légales en matière de protection des données personnelles
- ❖ Fournir des informations sur la source des données
- ❖ Utiliser des identifiants pérennes pour identifier les ressources

→ [Les principes de la qualité des données \(GBIF\)](#)

# SÉCURISER, STOCKER ET ARCHIVER LES DONNÉES



# 5. Sécuriser, stocker et archiver ses données

Quelles sont vos pratiques en matière de : stockage, sauvegarde, sécurité et archivage de vos données ?

- ❓ Quelles données pensez-vous supprimer en fin de projet ?  
Quelles sont les règles de suppression de ces données ?
- ❓ Quelles données doivent être archivées ? Pendant combien de temps ?  
Quelles sont les règles d'archivage des données ?
- ❓ Quelles données à conserver pour des raisons contractuelles, juridiques ?

# 5. Sécuriser, stocker et archiver ses données

Fonction fondamentale : la **conservation des données**

## Sauvegarde

**Copie des données** qui peut être utilisée pour restaurer les données originales dans le cas où ces dernières seraient endommagées ou perdues.

## Archivage

Il a pour finalité de **préserver des données anciennes / dormantes sur une longue durée** (> 1 an) dans l'éventualité d'un accès ultérieur (généralement obligation légale);

 Une archive est une **donnée originale** car transférée depuis son emplacement initial.

## Stockage

Répond a des problématiques d'**usage collaboratif** : dépôt, partage.

## 5. Sécuriser, stocker et archiver ses données

### → Systèmes de stockage des données

Support de stockage	Sécurité	Accès	Coût	Remarque d'utilisation
 <p>Ordinateur professionnel</p>	<p>★★★★☆</p> <p>Sujet au piratage informatique, aux détériorations et pannes</p>	<p>★★★★☆</p> <p>Pas adapté au partage, nécessite l'utilisation d'un support externe ou d'Internet (mail, cloud...)</p>	<p>★★★★★</p> <p>Pas de coût supplémentaire ou coût peu important</p>	<ul style="list-style-type: none"> <li>- Pour un stockage temporaire</li> <li>- Nécessité de crypter les données confidentielles et sensibles</li> </ul>
 <p>Support externe</p>	<p>★★★★☆</p> <ul style="list-style-type: none"> <li>- Sujet au vol, à la perte du support</li> <li>- Durée de vie limitée (dégradation du matériel)</li> </ul>	<p>★★★★☆</p> <p>Facilement transportable, il permet de transférer les données vers un autre ordinateur</p>	<p>★★★★★</p> <p>Pas de coût supplémentaire ou coût peu important</p>	<ul style="list-style-type: none"> <li>- Pour un stockage temporaire</li> <li>- Nécessité de crypter ou de sécuriser physiquement les données confidentielles et sensibles</li> </ul>
 <p>Serveur institutionnel</p>	<p>★★★★★</p> <p>Stockage fiable, durable et sécurisé (contre le vol, le piratage, les incendies...)</p>	<p>★★★★☆</p> <p>La connexion au serveur institutionnel ne facilite pas le travail avec des personnes extérieures</p>	<p>★★★★☆</p> <p>Coût assez important mais pas forcément répercuté sur l'utilisateur</p>	<ul style="list-style-type: none"> <li>- Pour un stockage plus pérenne</li> <li>- Adapté pour le stockage de données sensibles et des versions « stables » de vos données</li> <li>- Toutes les institutions ne proposent pas ce service</li> </ul>
 <p>Serveur Cloud</p>	<p>★★★★☆</p> <p>On ne sait pas vraiment où sont stockées les données, ni ce qu'elles deviennent</p>	<p>★★★★★</p> <p>Permet un travail synchronisé avec toutes les personnes ayant été autorisées au partage</p>	<p>★★★★☆</p> <p>Payant à partir d'une certaine limite de stockage</p>	<ul style="list-style-type: none"> <li>- Pour un partage avec des personnes externes à l'institution</li> <li>- Ne pas y mettre de données sensibles ou confidentielles</li> <li>- Pas de contrôle sur la procédure de sauvegarde des données</li> </ul>

## 5. Sécuriser, stocker et archiver ses données

### → Stockage sécurisé des données

Le stockage sécurisé est important pour assurer la continuité de l'exploitation des données sur du court terme.

### Sécurité physique

Pour ne pas perdre ses données, la règle simple de la sauvegarde : **3-2-1**

- ❖ Faire au moins **3 copies** de ses données
- ❖ Stocker les copies sur au moins **2 supports** différents
- ❖ Avoir au moins **1 sauvegarde hors site**



- ❖ **Zéro inquiétude** pour la sauvegarde des données

## 5. Sécuriser, stocker et archiver ses données

### → Stockage sécurisé des données

## Sécurité informatique

Pour se protéger des piratages et du vol de ses données, il est nécessaire :

- ❖ D'avoir un **antivirus à jour** pour minimiser les attaques informatiques



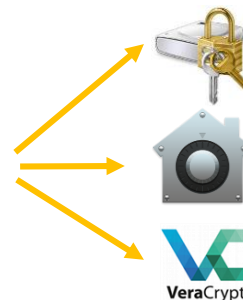
- ❖ De stocker ses données sur un ordinateur **non connecté au réseau**



- ❖ D'utiliser un **mot de passe** de niveau de protection **élevé** (longue chaîne de caractère avec majuscule(s), chiffre(s), caractères spéciaux)



- ❖ De **chiffrer** ses données



BitLocker sous Windows 7+

FileVault sous macOS

Solutions tierces

## 5. Sécuriser, stocker et archiver ses données

### → Recommandations pour le nommage des fichiers

Il est important d'avoir une stratégie de nommage des fichiers de données et de la respecter :

- ❖ Pour **récupérer et identifier plus facilement** les fichiers de données recherchés
- ❖ **Éviter les problèmes** lors de transfert et de partage
- ❖ Permettre leur **conservation à moyen et long terme**

Quelques bonnes pratiques de nommage des fichiers :

- ❖ Choisir un **nom court** (- de 100 caractères) **et significatif**
- ❖ Pas d'espace, pas d'accent, pas de caractères spéciaux
- ❖ Utiliser des abréviations standards (ex : CR pour compte rendu)
- ❖ Indiquer le **numéro de version du fichier**
- ❖ Indiquer le **nom de l'auteur**
- ❖ Renseigner la **date** et/ou l'heure
- ❖ Penser à la numérotation (saisir des **0 initiaux pour les tris**)

# 5. Sécuriser, stocker et archiver ses données

## Le Cirad propose...

Dataverse est ouvert comme lieu de stockage



→ <https://dataverse.cirad.fr/>

*Une formation Cirad aura lieu le 13 avril*

# BONNES PRATIQUES JURIDIQUES



# 6. Bonnes pratiques juridiques

## Cadre juridique de l'open data en France

### Open Data : politique d'ouverture des données :

par l'Etat et les collectivités publiques  fonds publics/ bien commun.

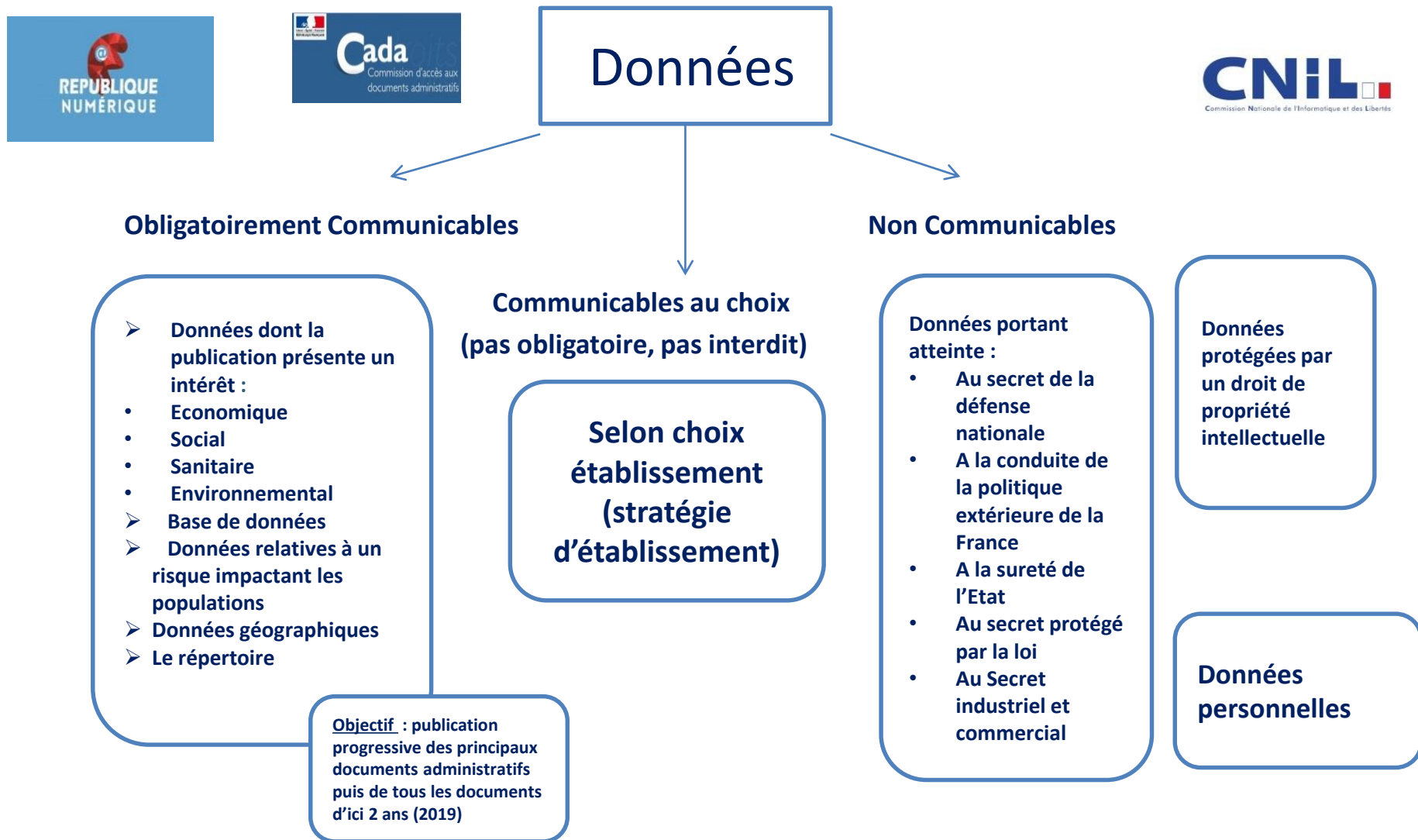
- ❖ Nombreux textes de référence (loi CADA, Loi Valter, Loi république Numérique, Code de la recherche, Code des relations entre le public et l'administration, Code de la Propriété Intellectuelle, directive Inspire...)
- ❖ Sur le plan Juridique l'OPEN DATA est une renonciation des droits sur les bases de données

### Evolutions : Loi République Numérique – 7 octobre 2016 :

- ❖ De la demande de communication à un **système de diffusion spontanée** ;
- ❖ Principe de **mise à disposition et de réutilisation gratuite** de l'information publique :
  - Disparition de l'exception des établissements de recherche ;
- ❖ Entre administrations, il ne peut pas y avoir de redevance pour se communiquer des données ;
- ❖ **Libre réutilisation des données/écrits scientifiques** publiées issues des activités de recherche financées au moins par moitié par l'argent public ;
- ❖ Autorisation du **Text et data mining**.

# 6. Bonnes pratiques juridiques

## Communication et diffusion



# 6. Bonnes pratiques juridiques

## Réutilisation des Données



Données  
communiquées/  
diffusées

**Principe :**  
**Libre réutilisation  
des données publiques  
(communication  
obligatoire)**

**Exception :**  
**Réutilisation des  
données soumise à  
autorisation**

### MINI

- ✓ **Pas de nécessité de licence :**
- ✓ Obligation de respecter les 4 conditions suivantes lors de la réutilisation
  - Non altération ;
  - Non dénaturation ;
  - Source ;
  - Date de dernière mise à jour.

### MAXI

- ✓ **Licence obligatoire :** fixée par décret/homologué et par l'Etat ;
  - Redevances ;
  - Restrictions liées à intérêt général et proportionnées ;
  - Ne pas restreindre concurrence ;
  - Utilisation commerciale ou non ;
  - .....

### Conditions :

- Si droit des tiers ;
  - Si droit de propriété intellectuelle
- Exception : le droit du producteur de BDD ne peut être opposé par l'administration.
- Si données produites dans mission SPIC, soumise à concurrence.

### Contexte :

- Mise en œuvre de la politique/stratégie d'établissement
- Respect des règles éthiques et déontologiques

# 6. Bonnes pratiques juridiques

## Open Data

- ❖ Dans quels cas dois-je diffuser mes données en open data ?
- ❖ Si la loi me l'impose : loi Cada, loi Valter, loi pour une république numérique (Dite Loi Axelle Lemaire)
- ❖ Si le bailleur me l'impose : condition prévue dans les règles de participation
- ❖ Si le consortium me l'impose : les parties peuvent décider collectivement de diffuser les données en open data à l'issue du projet.

# 6. Bonnes pratiques juridiques

## Focus sur les données personnelles

- ❖ Cadre réglementaire : Règlement Général sur la protection des données personnelles (RGPD) qui entre en vigueur en France le 25 mai 2018

- ❖ Qu'est-ce qu'une donnée personnelle ?

*Toute information se rapportant à une personne physique identifiée ou identifiable ; est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale.*

# 6. Bonnes pratiques juridiques

## Introduction de la notion de finalité

- ❖ Le RGPD introduit la notion de **finalités d'usage des données personnelles**.
- ❖ Les données personnelles sont collectées pour des finalités déterminées, explicites et légitimes, et ne peuvent pas être traitées ultérieurement d'une manière incompatible avec ces finalités ; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré comme incompatible avec les finalités initiales (limitation des finalités).

# 6. Bonnes pratiques juridiques

## Focus RGPD

OPEN DATA =  
LIBRE REUTILISATION  
DES DONNEES

DONNEES  
PERSONNELLES =  
PRINCIPE DE  
FINALITE DE  
TRAITEMENT

ANONYMISATION  
=  
OUVERTURE  
POSSIBLE

# 6. Bonnes pratiques juridiques

## Focus sur les données sensibles

Qu'est-ce qu'une donnée sensible ?

Les données sensibles sont celles qui font apparaître, directement ou indirectement, les **origines raciales ou ethniques**, les **opinions politiques**, **philosophiques ou religieuses** ou **l'appartenance syndicale** des personnes, ou sont relatives à la **santé** ou à la **vie sexuelle** de celles-ci.

Par principe la collecte et le traitement de ce genre de données sont interdits.

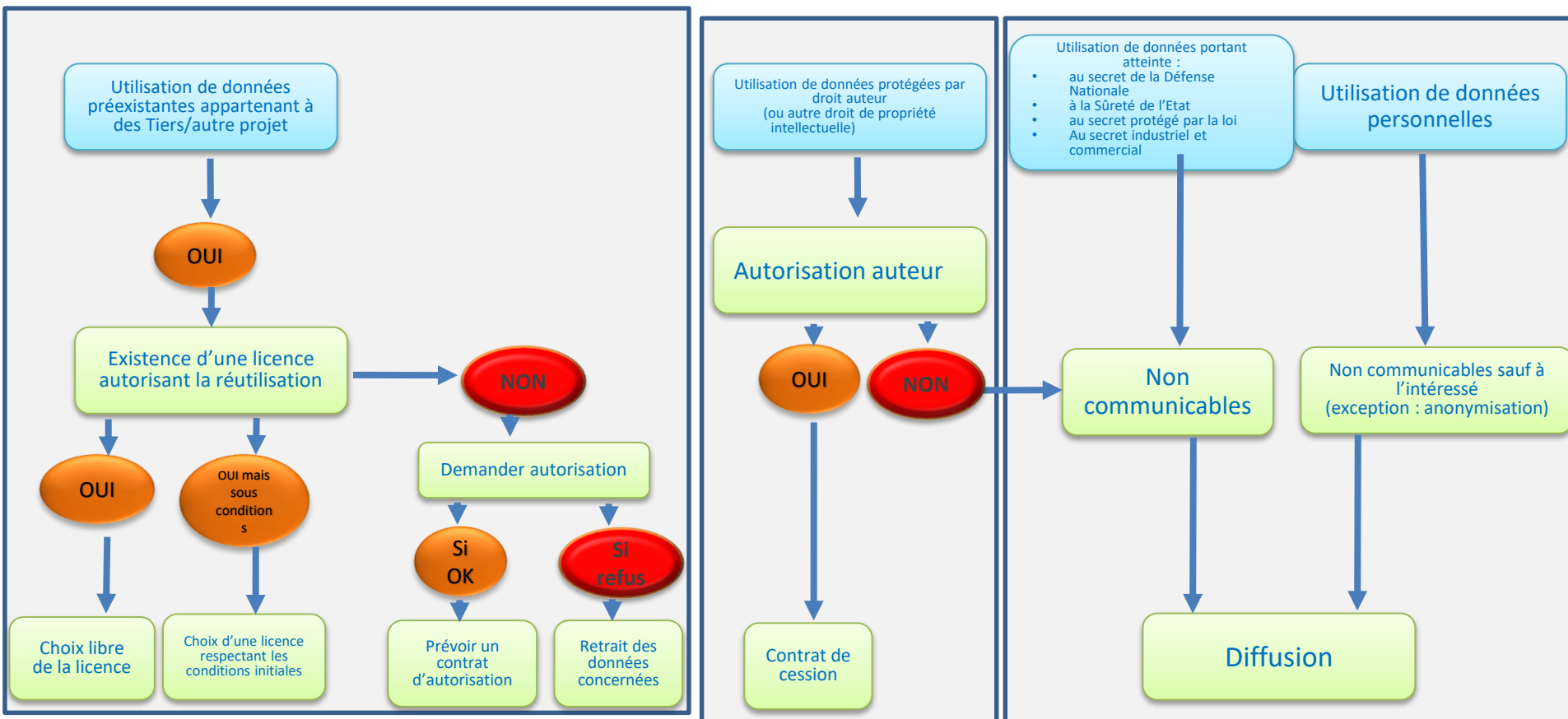
Cependant dans la mesure où la finalité du traitement l'exige, ne sont pas soumis à cette interdiction :

- ❖ Les traitements pour lesquels la personne concernée a donné son consentement exprès
- ❖ Les traitements justifiés par un intérêt public après autorisation de la CNIL ou décret en Conseil d'État
- ❖ La collecte et le traitement de ces données doivent dans ces hypothèses être justifiées au cas par cas au regard des objectifs poursuivis



# 6. Bonnes pratiques juridiques

## Données générées par le projet



# 6. Bonnes pratiques juridiques

## La licence – Outil de diffusion

- ❖ La licence est un contrat qui permet de déterminer les conditions de diffusion et de réutilisation des jeux de données/bases de données
- ❖ La licence peut permettre un simple droit d'accès pour consultation ou autoriser l'extraction des données
- ❖ La licence peut être ouverte (ex : licence Creative commons, ODBL...) ou fermée (licence propriétaire)
- ❖ Lors de la diffusion de la base de donnée, je m'interroge sur la licence sous laquelle je diffuse

# 6. Bonnes pratiques juridiques

## La licence Creative Commons



Licence CC	Bouton	Explications	\$ Exploitation permise?	♻️ Remix permis?
<b>Attribution</b>		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre est permise.</li> <li>Vous pouvez remixer l'oeuvre.</li> <li>Le partage de l'oeuvre est permis.</li> </ul>		
<b>Share Alike</b> Partage à l'identique		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre est permise.</li> <li>Vous pouvez remixer l'oeuvre.</li> <li>Cette licence doit toujours être utilisée sur toutes vos versions dérivées de l'oeuvre originale.</li> <li>Le partage est permis.</li> </ul>		
<b>No Derivative</b> Modification non permise		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre est permise.</li> <li>Vous <b>NE</b> pouvez <b>PAS</b> remixer l'oeuvre.</li> <li>Le partage est permis.</li> </ul>		
<b>Non-Commercial</b> Usage commercial non permis		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre <b>n'est PAS</b> permise.</li> <li>Vous pouvez remixer l'oeuvre.</li> <li>Le partage est permis.</li> </ul>		
<b>Non-Commercial</b> Usage commercial non permis + <b>Share Alike</b> Partage à l'identique		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre <b>n'est PAS</b> permise.</li> <li>Vous pouvez remixer l'oeuvre.</li> <li>Cette licence doit toujours être utilisée sur toutes vos versions dérivées de l'oeuvre originale.</li> <li>Le partage est permis.</li> </ul>		
<b>Non-Commercial</b> Usage commercial non permis + <b>No Derivative</b> Modification non permise		<ul style="list-style-type: none"> <li>On doit citer <b>QUI</b> est l'auteur de l'oeuvre originale.</li> <li>L'utilisation commerciale de l'oeuvre <b>n'est PAS</b> permise.</li> <li>Vous <b>NE</b> pouvez <b>PAS</b> remixer l'oeuvre.</li> <li>Le partage est permis.</li> </ul>		

# 6. Bonnes pratiques juridiques

## Outils d'aide au choix de licence

- ❖ <http://licentia.inria.fr/licenseservice>
- ❖ <https://ufal.github.io/public-license-selector/>
- ❖ <https://eudat.eu/services/userdoc/license-selector>

# 6. Bonnes pratiques juridiques

## Exemples de licences ouvertes

### ❖ Creative Commons

❖ **ODBL** : Open Database License (ODbL) est un contrat favorisant la libre circulation des données.

- Licence autorisant l'exploitation commerciale
- Partage à l'identique des conditions initiales

❖ **Etalab** (FR) : facilite et encourage la réutilisation des données publiques mises à disposition gratuitement

- Une licence ouverte, libre et gratuite ;
- Une licence qui promeut la réutilisation la plus large en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données ;
- Une licence qui s'inscrit dans un contexte international en étant compatible avec les standards des licences Open Data développées à l'étranger (Open Government Licence, ODC-BY, CC-BY 2.0...).

# **PARTAGE ET VALORISATION DES DONNÉES**

# 7. Partage et valorisation des données



Pendant et après le projet

- ❖ Quels jeux de données ?
- ❖ Quand ?
- ❖ Où ? → quel entrepôt de données ?
- ❖ Comment ? → quelles modalités de partage ?
- ❖ Utiles pour qui ?
- ❖ Quelles réutilisations possibles ?
- ❖ Quelles perspectives d'application ou de développement ?

# 7. Partage et valorisation des données



Pendant le projet

## ❖ Avec qui ?

- Tous les partenaires du projet
- D'autres personnes / partenaires hors projet

## ❖ Où ?

- Alfresco (documents)
- Dataverse Cirad (données)

## ❖ Comment ? Modalités de partage et d'accès en interne

- Les jeux de données seront déposés sur le Dataverse du Cirad et partagés avec tous les partenaires du projet.
- Présence de partenaires privés dans le consortium → détails dans l'accord de consortium



# 7. Partage et valorisation des données



Après le projet

- ❖ Quels jeux de données ?
- ❖ Quand ?
- ❖ Où ? → quel entrepôt de données ?
- ❖ Comment ? → quelles modalités de partage ?

# 7. Partage et valorisation des données



Partage fortement recommandé  
(*institution, bailleurs, revues, ...*)

## ❖ Choix des jeux de données à partager



*As open as possible, as closed as necessary*

protection des données sensibles, personnelles, de santé,  
partenariats privés,... → à justifier dans le PGD

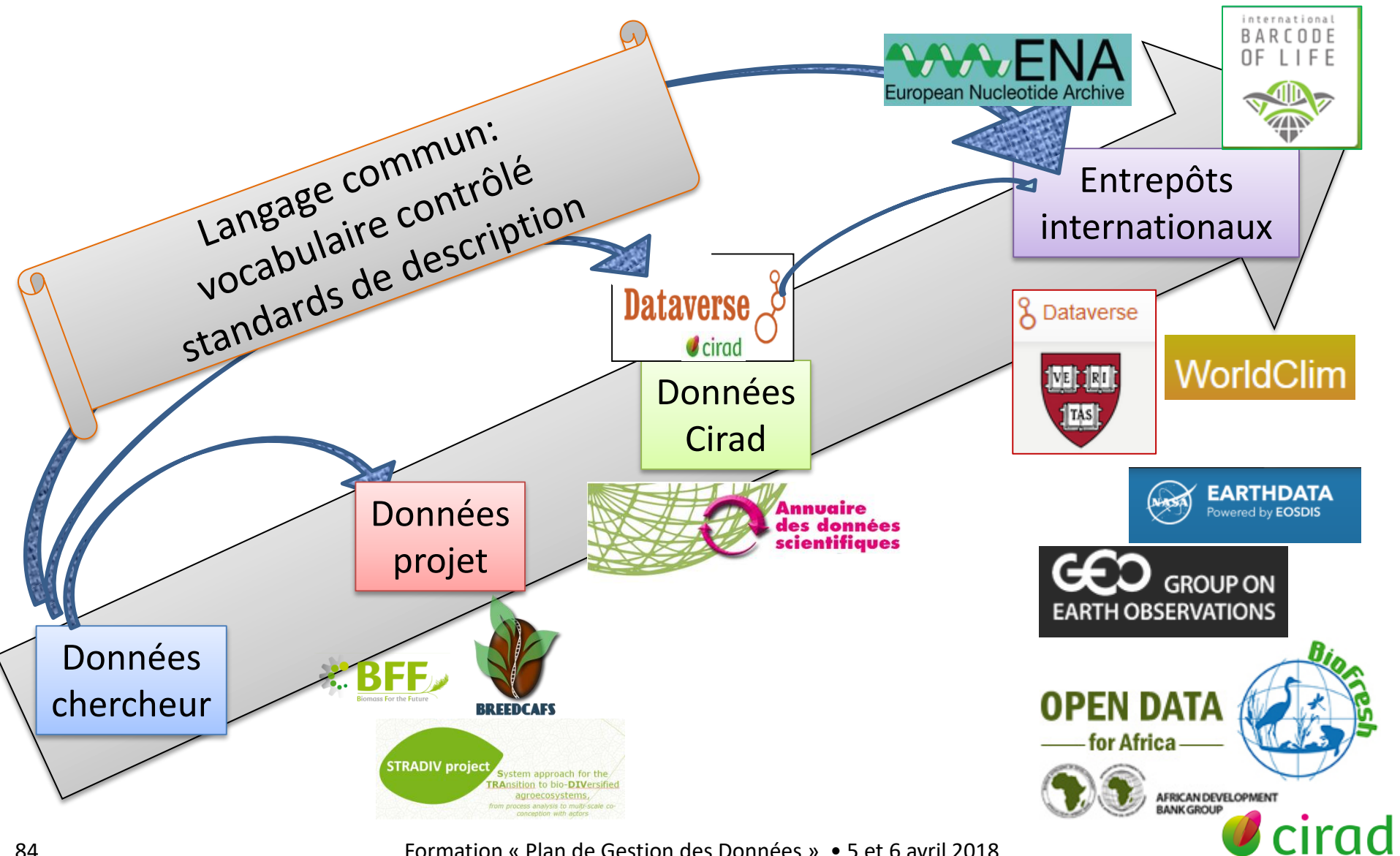
## ❖ Choix des modalités d'accès à vos données

licence ouverte/accès restreint/enregistrement

période d'embargo

limiter réutilisation (non commerciale, selon projet)

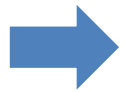
# 7. Partage et valorisation des données



## 7. Partage et valorisation des données

### → Quels jeux de données partager ? 1/2

- ❖ Données qui feront l'objet d'articles de recherche
- ❖ Données ayant un potentiel de réutilisation
  - Nouvelles analyses, nouvelles questions de recherche
  - Méta-analyses,
  - Changement d'échelle : spatiale, temporelle, analyses stat.
  - Paramétrer ou nourrir des modèles,
  - Développement commercial, de services...
  - etc.
- ❖ Données utiles (jeux de données contrôles, lots témoins)
- ❖ Données dont vous savez qu'elles présentent un intérêt pour certains publics et que vous n'utiliserez plus

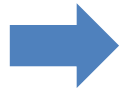


#### Cas particuliers, à justifier dans le PGD

- ❖ Données issues de partenariats (privés, CGIAR, etc.)
- ❖ Données sensibles :
  - espèces protégées ou envahissantes,
  - données cliniques, issues d'expérimentations animales
  - données économiques, personnelles
  - issues de ressources biologiques du Sud (APA), etc.
- ❖ Données stratégiques que vous souhaitez exploiter :
  - identification de marqueurs génétiques, d'aromes
  - création d'une appli, d'une base de données originale
- ❖ Jeux de données contenant des données préexistantes (produites par d'autres : licences à vérifier)

## 7. Partage et valorisation des données

### → Identifier ses futurs utilisateurs



Quels publics seraient intéressés ?

- ❖ Communautés scientifiques
- ❖ Enseignants
- ❖ Décideurs
- ❖ Secteur privé, créateurs de start-up
- ❖ Grand public
- ❖ ONGs, associations internationales influentes
- ❖ Journalistes

## 7. Partage et valorisation des données

### → Evaluer la valeur scientifique de vos données

#### ❖ Données rares ou uniques

- uniques et impossible à répéter
- groupes difficilement accessibles
- phénomènes rares

#### ❖ Données à forte valeur scientifique

- données de référence
- reproduction difficile ou coûteuse
- intérêt pour des pays du Sud

#### ❖ Données ayant une valeur économique

- perspectives d'application, de développement commercial

#### ❖ Données présentant un intérêt pour la société

#### ❖ Données ayant une valeur environnementale

- contribuant à un thème majeur:

ex: changement climatique, gestion ressources naturelles, santé publique

# 7. Partage et valorisation des données



## Potentiel de réutilisation dans le PGD

- ❖ S'il existe déjà des données similaires, quel est l'apport de ce jeu de données ?
- ❖ Quelles réutilisations possibles ?
- ❖ Quelles perspectives d'application ou de développement ?
  - dans votre domaine
  - dans d'autres domaines
- ❖ Option: recherche de nouvelles collaborations

*“Dr Judy Ford would be happy to be involved in future collaborations that involve these data....”*



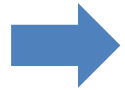
## 7. Partage et valorisation des données

### → Quand partager vos données ?


- ❖ Vous avez exploité vos données
- ❖ Vous avez publié vos résultats de recherche
- ❖ Vous avez mis en forme vos données et métadonnées
- ❖ Vos données ont été anonymisées
- ❖ Vos partenaires sont d'accord
- ❖ Période d'embargo possible
- ❖ Accès restreint ou accès sur demande

## 7. Partage et valorisation des données

### → Où partager vos données ?



Dans un entrepôt de données  
→ optimise les possibilités de réutilisation

- ❖ Entrepôt adapté à vos données  
thématique, multidisciplinaire, institutionnel  
reconnu (notoriété) dans votre discipline
- ❖ Correspondant aux demandes : bailleur / institution /  
partenaires / revue où vous publiez
- ❖ Correspondant aux publics scientifiques visés
- ❖ Acceptant les modalités de diffusion que vous souhaitez  
 qqs entrepôts en CC0
- ❖ Qui délivre un identifiant numérique pérenne et unique

## 7. Partage et valorisation des données

### → Standards de métadonnées disciplinaires



Connaissez-vous des **entrepôts de données** dans  
**votre domaine ?**

pour trouver des données fiables ?

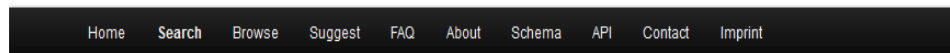
pour déposer des données ?

# 7. Partage et valorisation des données

## → Les entrepôts de données



→ 2000 entrepôts



Search for Repositories (1335 reviewed repositories)

Search bar with a magnifying glass icon and the word "Search".

Filters:

- Subject: Add subjects
- Content Type: Add content types
- Country (of the responsible institutions): Add countries

Checkboxes:

- ☐ Certificates
- ☐ Open Access
- ☐ Persistent Identifier

Reset filter button.

1335 results (1 - 25)

Page numbers: 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, ..., 54

Sort by: Weight

### Liste d'entrepôts de données (*Data repositories*)

dans les thèmes du Cirad

Laurence Dedieu, janvier 2018

### Table des matières

Qu'est-ce qu'un entrepôt de données ?	2
Méthode et critères de sélection des entrepôts de données dans ce document	3
Entrepôts généralistes	3
Entrepôts de génétique, génomique et biologie moléculaire	7
Entrepôts de séquences de protéines et protéomique	12
Entrepôts de données en métabolomique	14
Entrepôts de données sur les interactions moléculaires	15
Entrepôts de données de structure moléculaire	16
Entrepôts de données en taxonomie et phylogénie	24
Entrepôts de données en génétique animale	26
Entrepôts de données en biologie et écologie animale	27
Entrepôts de données en entomologie	28
Entrepôts de données en nématologie	30
Entrepôts de données en microbiologie	31
Entrepôts de données sur des interactions hôte/pathogènes	34
Entrepôts de données en mycologie	35
Entrepôts en génétique et biologie des plantes	19
Entrepôts de données en agriculture et foresterie	37
Entrepôts de données en écologie, environnement et biodiversité	39
Entrepôts de données en sciences de la terre	42
Entrepôts en données d'analyse du cycle de vie (ACV)	45
Entrepôts d'images	46
Entrepôts de modèles	49
Entrepôts de données en Sciences Humaines et Sociales	50
Entrepôts de données en économie	53
Entrepôts nationaux de données en SHS	54
Entrepôts de données situés en Afrique	58
Les licences de libre diffusion des données scientifiques	60
Liens utiles et références	61

## 7. Partage et valorisation des données

### → Les entrepôts de données

❖ **Institutionnels** Dataverse

❖ **Europe** Zenodo, B2Share

❖ **Généralistes** Figshare, Dryad

❖ **Editeurs** BMC (GigaDB) ; Ubiquity Press (Dataverse)

❖ **Thématiques**

- GenBank/ENA, GEO, Barcode of Life Data Systems, ArrayExpress
- UniProt, PeptideAtlas, MetaboLights, Intact
- Ensembl Plants, Gramene, Plant Metabolic Network, TropGeneDB
- Movebank, WormBase, ViPR, MycoBank, ComBase, FLOW
- Pangaea, SEANOE
- Dataverse, ICPSR, DataFirst, Quetelet, beQuali

# 7. Partage et valorisation des données

## → Où partager vos données ?



- ❖ Dépôt des données dans l'entrepôt du Cirad :



- ❖ Duplication de quelques métadonnées dans un entrepôt thématique partagé par votre communauté scientifique, avec le lien vers les données sur le Dataverse



## 7. Partage et valorisation des données

### → Valoriser vos données



Faites savoir que vos données sont de qualité  
qu'elles ont un potentiel de réutilisation  
qu'elles sont disponibles

- ❖ Publier un **datapaper**
- ❖ Publier un **article de recherche**
- ❖ Rédiger une brève pour un **magazine** spécialisé
- ❖ Contribuer à un **blog**, ....



Open Data Journal for Agricultural Research

# 7. Partage et valorisation des données

## → Le PGD peut être publié

Pensoft



Research Ideas and Outcomes  
The Open Science Journal

<https://riojournal.com/>

Data Management Plan: Empowering Indigenous Peoples and Knowledge Systems Related to Climate Change and Intellectual Property Rights

Cath Traynor

Data Management Plan doi: 10.3897/rio.3.e15111

25-07-2017 Unique: 296 | Total: 476 Reprint: € 2,30

See collection

HTML XML PDF

Data Management Plan: Opening access to economic data to prevent tobacco related diseases in Africa

Lynn Woolfrey

Data Management Plan doi: 10.3897/rio.3.e14837

24-07-2017 Unique: 242 | Total: 392 Reprint: € 2,60

See collection

HTML XML PDF

Data Management Plan: HarassMap

Reem Wael

Data Management Plan doi: 10.3897/rio.3.e15133

18-07-2017 Unique: 239 | Total: 374 Reprint: € 2,60

See collection

HTML XML PDF

Data Management Plan: IDRC Data Sharing Pilot Project

Cameron Neylon

Data Management Plan doi: 10.3897/rio.3.e14672

27-06-2017 Unique: 320 | Total: 537 Reprint: € 2,60

See collection



HTML XML PDF

Data Management Plan: Brazil's Virtual Herbarium

Dora Ann Lange Canhos

Data Management Plan doi: 10.3897/rio.3.e14675

27-06-2017 Unique: 360 | Total: 590 Reprint: € 2,60

See collection

HTML XML PDF



Deliverable 1.1

Data Management Plan

DISSEMINATION LEVEL		
PU	Public	X
CO	Confidential, only for members of the consortium (including the Commission Services)	



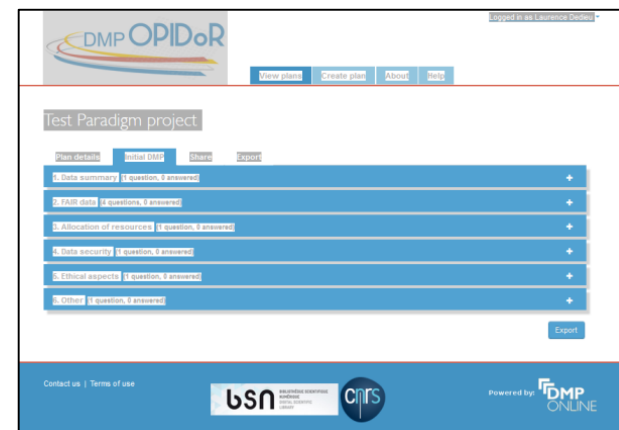


**CONCLUSION**

# Le Plan de Gestion des Données

- ❖ Décrit **tous les jeux de données** produits dans un projet  
Y compris ceux n'ayant pas vocation à être partagés  
et leur gestion pendant et après le projet
- ❖ Incite à utiliser des **méthodes/protocoles, normes/métadonnées reconnus dans la discipline**
- ❖ Document **évolutif** : V1 (6 mois), V2 (18 mois), V3 (fin)  
La V1 peut ne pas traiter toutes les questions
- ❖ À préparer **le plus tôt possible**
- ❖ **Outil d'animation** dans un collectif,  
doit être partagé entre partenaires

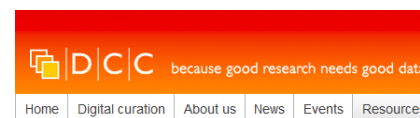
# Pour vous accompagner



<https://intranet-dist.cirad.fr/>



<https://coop-ist.cirad.fr/>



Search by Discipline



Biology



Earth Science



General Rese



Physical Science



Social Science & Humanities



**FIN**